Restricted maximum likelihood from symmetry.

Inge S. Helland *

Abstract

If a natural non-transitive group is attached to a statistical model, minimum risk equivariant estimators could be used on orbits, and for the orbit index, maximum likelihood estimation from the sample orbit index. This is used to motivate REML.

KEY WORDS: Group; Symmetry; Maximal Invariant; Maximum Likelihood Estimation; Mixed Linear Models; Orbit Index; Orbits; REML; Restricted Maximum Likelihood; Variance Components.

1 Basic approach.

Maximum likelihood is the default estimation method in most statistical applications, even though it is well known that it can be motivated properly only for large data sets. A somewhat more specific statement is to say that maximum likelihood requires the number n of data points to be large compared to the number p of parameters in the model. If this condition does not hold, we have basically only two general alternative methods of estimation (Lehmann & Casella, 1998). The first alternative is to try to find the best unbiased estimator, most commonly through conditioning

*Department of Mathematics, University of Oslo, P.O.Box 1053 Blindern, N-0316 Oslo, Norway. E-mail: ingeh@math.uio.no

and the Rao-Blackwell theorem. The second alternative is to impose some symmetry on the situation and try to find the best equivariant estimator. We will use the second approach here.

Assume that a group G on the sample space has been chosen. In general the choice of group is crucial; different groups typically lead to different estimators, but this is really not more strange than the fact that different loss functions, or for that sake different variants of the same model, lead to different estimators. What is important, is to try to choose a group which is felt to express a natural symmetry of the situation.

From G, a group $\bar{G}$ on the parameter space is induced by

$$P_{\bar{g}\theta}(y \in A) = P_\theta(gy \in A),$$

the basic requirement being that the model is closed under this group and that the loss function is invariant.

Call a parametric function ψ invariantly estimable if $\psi(g\theta)$ always is a function of $\psi(\theta)$; then a new group G^* is induced on the image of ψ. An estimator $\hat{\psi}$ is called equivariant if $g^*\hat{\psi}(x) = \hat{\psi}(gx)$ always.

The simplest situation is when the group $\bar{G}$ is transitive on the parameter space. (This means that for any fixed θ_0, we have that $\bar{g}\theta_0$ runs through the whole parameter space when $\bar{g}$ runs through $\bar{G}$.) Then the risk function (expected loss) will be a constant for all θ when the estimator is fixed, so the risk is a uniquely determined function of the estimator. This strongly suggests that there always exists a unique best equivariant estimator, which in fact holds quite generally. When $\bar{G}$ is not transitive, a similar uniqueness property holds on the orbits of the group, where an

orbit is defined as the collection of parameter values $\bar{g}\theta_0$, where $\bar{g}$ runs through $\bar{G}$ and θ_0 is fixed. The following facts are well known:

1. The orbit index τ for a non-transitive group in the parameter space is the same as the maximal invariant under $\bar{G}$.

2. A similar statement holds in the sample space. The orbit index a here has a distribution which only depends upon τ.

3. The risk function is constant on orbits in the parameter space.

The last statement again strongly suggests that estimation of the parameter on the orbit leads to a unique best equivariant solution, which indeed is the case quite generally; more explicitly, the estimator can be expressed as an integral of the Pitman type.

But after this, the orbit index (maximal invariant) must be estimated in other ways, and a natural solution is to use maximum likelihood *using the orbit index in the sample space.*. The main purpose of this note is to point out that the restricted maximum likelihood (REML) estimator of dispersion parameters in linear mixed models is just a straightforward example of this.

The REML estimator was first proposed (for balanced data) by W.A. Thompson (1962), and then it was independently proposed and applied for unbalanced data by Patterson & R. Thompson (1971). After having competed with some other variance component estimators for several years, it is now the standard procedure, with a considerable attached literature.

I don't believe that the result given below is very original in a strict meaning of this term. For instance, it is hinted at on p. 191 in Lehmann & Casella (1998). Nevertheless, I feel that it is useful to give an explicit discussion in a grouptheoretical

setting, both since this perspective seems to be enlightening, and since it gives the possiblity to extend the estimation principle to many other situations. None of the 94 references in the encyclopedia article on REML by Speed (1997) seem to have been written from this perspective. The only explicitely related paper which I know of, is McCullagh (1996), who uses a more abstract approach. A completely different motivation for REML is given in Smyth & Verbyla (1996), where further references to the question of motivation are given.

2 The model.

Let the vector of observations y be modelled as

$$\mathrm{N}_n(X\beta, \Sigma), \tag{1}$$

where X is a known $n \times p$ matrix of rank p, and where Σ depends upon r unknown parameters γ. We will assume that γ varies over some open set, and that otherwise conditions (Lehmann & Casella, 1998) for the existence and uniqueness of the solutions of the likelihood equations corresponding to the nonsingular distributions given below, are satisfied.

An important special case of this model is given by the mixed model

$$y = X\beta + \sum_{k=1}^{r} Z_k u_k,$$

where the matrices Z_k have dimensions $n \times q_k$ and rank q_k, and where the u_k's are independent with $u_k \sim \mathrm{N}_{q_k}(0, \sigma_k^2 I)$.

The parameters of the model are $\theta = (\beta, \gamma)$, and in the mixed model case $\gamma = (\sigma_1^2, \ldots, \sigma_r^2)$ are called the variance components.

This model again contains as special cases all common linear models, in particular those of analysis of variance, balanced or unbalanced. When $r > 1$, a non-trivial set of variance components is to be estimated.

3 The group.

The expectation part of the model (1) simply means that $\mathrm{E}(y)$ is assumed to belong to the p-dimensional known vector space $V = \mathrm{span}(X)$. A natural symmetry group attached to the model is therefore the group G of translations in this space. As a first observation, if $y \to gy = y + c$ for some fixed $c \in V$, and if the model holds for y, then it also holds for gy. Hence the model is invariant under this group.

The induced group $\bar{G}$ in the parameter space is given by $(\beta, \gamma) \to (\beta + b, \gamma)$, where $c = Xb$. Note that c runs through V if and only if b runs through all of R^p.

Obviously, $\bar{G}$ is not transitive, and the orbits are just indexed by γ.

4 Estimation within orbits.

For a fixed orbit, we have that $\Sigma = \Sigma(\gamma)$ is fixed, and the maximum likelihood estimate is

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y.$$

Extending the discussion of Lehmann & Casella (1998), Section 3.4, we can also show that this estimator gives the minimum risk equivariant estimator of linear combinations $\phi = a'\beta$ for convex and even loss functions. Of course, this is also the

weighted least squares estimator, which is well known to have good properties and is much used in many areas of applied statistics. The only difficulty in the present setting is that it depends upon unknown parameters, the indices of orbits.

5 Estimation of orbit indices.

This is the crucial point, and the argument relies on an easily proved property of groups: The maximal invariant in sample space has a distribution which depends only upon the orbit index (maximal invariant) in the parameter space. Hence this gives a natural setting for estimation.

Theorem 1.

(a) In the model (1) under the group G, the maximal invariant can be expressed as

$$r = (I - X(X'X)^{-1}X')y.$$

(b) Let A be an $n \times (n-p)$ matrix of full rank $n-p$ such that $A'X = 0$. Then an equivalent orbit index is given by $z = A'r = A'y$. This variable z will have a non-singular distribution.

(c) The maximum likelihood estimator of γ found from the distribution of z is independent of the choice of the matrix A with the stated properties.

Remark.

The maximum likelihood estimator referred to in Theorem 1 will be the REML estimator for models of the form (1), in particular for mixed linear models. It is obvious from the setting that it will give an estimator of just the orbit index γ, and it is also obvious that this estimator will have many of the ordinary properties of maximum likelihood estimators. It should also be quite clear that as a principle of estimation this can be generalized to many other situations where a natural group can be associated to the statistical model.

Proof.

(a) It is clear that if $y \to y + c$, where $c \in V = \text{span}(X)$, then $r \to r$, so r is invariant. Since y can be recovered from r and the projection of y upon V, and since no part of this projection can be invariant under translations in V, we must have that r is maximal invariant.

(b) From the model equation of the form $y = X\beta + e$, where $e \sim \mathrm{N}_n(0, \Sigma)$, we see that $r = Pe$, where $P = I - X(X'X)^{-1}X'$ is the projection upon the $(n-p)$-dimensional space orthogonal to V. From this we see directly that $z = A'e$ has a distribution which is independent of β, specifically, $\mathrm{N}_{n-p}(0, A'\Sigma A)$, which is non-singular, since the covariance matrix must have rank $n - p$.

(c) For any $n \times (n-p)$ matrix B of rank $n - p$ such that $B'X = 0$ we must have that the coloumns of B must span the space orthogonal to V; hence $B = AC$ for a non-singular matrix C. This implies that $B'y = C'A'y$, and the likelihoods of $A'y$

and $B'y$ can be simply transformed into each other.

6 Calculation.

In a general estimation procedure, it may be impractical to find an unspecified matrix A whose columns span the space orthogonal to $V = \text{span}(X)$, so practical computation algorithms use other techniques. One of the early papers on computation in REML is Corbeil & Searle (1976), while a modern survey and many references can be found in Speed (1997). The methods are much used by animal breeders, and this community has also done an important job in constructing efficient progams for REML estimation in large data sets.

7 A simple example.

Here is the simplest possible example: Let $y_1, \ldots, y_n$ be independently $\mathrm{N}(\mu, \sigma^2)$. The REML estimator of σ^2 is, according to the receipt given above, found as follows: Take $1 = (1, \ldots, 1)'$. and let A be any $n \times (n-1)$ matrix of full rank satisfying $A'1 = 0$. Then a simple calculation gives that the maximum likelihood estimator from $z = A'(y - \bar{y}1) = A'y$ is given by

$$\hat{\sigma}^2 = \frac{1}{n-1} z'(A'A)^{-1} z = \frac{1}{n-1} y'A(A'A)^{-1}A'y.$$

Here it also can be seen directly that the estimator is independent of the choice of A, since the resulting projection equals

$$A(A'A)^{-1}A' = I - 11'/n.$$

Inserting this gives the ordinary variance estimator wich is unbiased, or, more important, has a denominator with the correct degrees of freedom.

Consider now a possibly unbalanced one-way analysis of variance situation, i.e., k independent groups, where group j contains n_j independent observations, each $\mathrm{N}(\mu_j, \sigma^2)$. The REML estimator for σ^2 can here be derived from a simple extension of the result of the previous paragraph, and will be

$$\hat{\sigma}^2 = \frac{1}{\sum n_j - k} \sum_j \sum_i (y_{ji} - \bar{y}_{j\cdot})^2.$$

The corresponding maximum likelihood estimator is biased, and is found by deleting the $' - k'$ in the denominator here. The bias can be considerable if k is large and the numbers of observations in the groups are small. For instance, if $n_1 = \ldots = n_k = 2$, then the denominator in the REML estimator is k, as it should be, while the maximum likelihood estimator is too small, with a denominator $2k$. A similar example from block experiments can be traced back to Neyman & Scott (1948).

8 Concluding remarks.

The REML principle has turned out to be very useful, for instance in animal breeding, but also in other cases where linear mixed models are used. According to Speed (1997), in the 1970s REML was simply one of a number of methods of estimating dispersion parameters, but now it is becoming *the* preferred method. By what we hope to have illustrated here using very simple arguments, REML estimation can be regarded as an instance of a general symmetry based estimation principle which seems to have the potential for even further applications.

References.

Corbeil, R.R. & S.R. Searle, 1976. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, **18**, 31-38.

Lehmann, E.L. & G. Casella, 1998. *Theory of Point Estimation.* 2. edition. Springer-Verlag, New York.

McCullagh, P., 1996. Linear models, vector spaces, and residual likelihood. In: *Modelling Longitudinal and Spatially Correlated Data.* Springer Lecture Notes No. 122, 1-10.

Neyman, J. & E.L. Scott, 1948. Consistent estimators based on partially consistent observations. *Econometrica*, **16**, 1-32.

Patterson, D. & R. Thompson, 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545-554.

Smyth, G.K. & A.P. Verbyla, 1996. A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *J. R. Statist. Soc.* B, **58**, 565-572.

Speed, T., 1997. Restricted maximum likelihood (REML). In: *Encyclopedia of Statistical Sciences.* Update volume 1, Wiley, New York

Thompson, W.A., Jr., 1962. The problem of negative estimates of variance components. *Ann. Math. Statist.*, **33**, 273-289.